Wednesday Afternoon

Speech Results on Resource Management Task

David S. Pallett
National Institute of Standards and Technology

The session on "Resource Management Task Speech Recognition Results" included presentations discussing progress in speech recognition using the DARPA Resource Management Speech Database. Benchmark Test results were presented by DARPA contractors at BBN, CMU, MIT Lincoln Laboratory, MIT Laboratory for Computer Science, and SRI. Summary tabulations of a portion of these test results are presented in this report. Additional preliminary results using this database were presented by two other organizations: AT&T Bell Laboratories and IBM Watson Research Laboratories.

OVERVIEW

Two sets of Benchmark Test material were selected from the Resource Management Database prior to this meeting: one set of test material for speaker dependent systems included 25 sentence utterances from each of the 12 speakers in the Speaker Dependent Evaluation ("tdde") subset, and the other set for speaker independent systems included 30 utterances from each of 10 speakers in the Speaker Independent ("tdie") subset. The speaker dependent test material was used by BBN and MIT Lincoln Laboratory for their speaker dependent systems. The speaker independent test material was used for tests of speaker independent systems at AT&T Bell Laboratories, CMU, MIT Laboratory for Computer Science, MIT Lincoln Laboratory, and SRI.

Schwartz, from BBN, noted that the speaker-dependent hidden Markov model (HMM) system used at BBN for their Benchmark Tests was based on studies that included a comparison of methods for smoothing discrete probability functions. Reference (3) describes experiments conducted prior to the DARPA Benchmark Tests that indicate the best method (in those experiments) was triphone co-occurrence smoothing, a method based on deriving a probabilistic co-occurrence matrix between different vector-quantized spectra.

Paul reported Benchmark Test results at MIT Lincoln Laboratory (2) for the "Lincoln stress resistant HMM" continuous speech recognition system for both speaker-dependent and speaker-independent tasks, using both Benchmark Test sets. The system used for results reported earlier, the "June 88" system, was a "continuous observation density HMM with triphone (left and right context sensitive phone) models. For this system, word- context-free [WCF] triphones (i.e., the triphone contexts included word boundaries, but excluded the phone on the other side of the [word] boundary) were used.

More recently, word-context-dependent models were also trained, providing the "recognizer with a set of models for [both] the observed word boundaries and a set of WCF models to be used for word boundaries allowed by the grammar but not observed in the training data. This reduces the number of phones extrapolated by the recognizer." Paul's results show that the speaker-dependent system was improved significantly by the addition of the word boundary triphone model. However, the speaker-independent system results were worse than for the WCF system, and "the word-context-dependent system appears to be too detailed a model for the available speaker independent training data".

Paul also reported results using both test sets for his speaker-independent system, trained on 109 speakers (3990 sentence utterances), to provide a comparison between the two test sets. For the case of the word-pair grammar, the word error was 11.4% for the speaker-dependent test set, and 9.8% for the speaker-independent test set, possibly suggesting that the speaker-independent test set may be somewhat easier to recognize than the speaker-dependent test set. Finally, Paul also provided comparisons of performance with speaker-independent systems trained on both 72 speakers (2880 sentence utterances) [the LL (72) system of Table 2] and 109 speakers [LL (109) in Tables 1 and 2], with better results occurring for the better-trained systems.

Recent improvements in the CMU SPHINX speech recognition system were described by Lee (1). These enhancements include function-phrase modeling, between-word coarticulation modeling and corrective training. "Function word/phrase dependent models" have been incorporated to more explicitly model "the most difficult vocabulary", involving a set of 42 function words and each of the 105 phones contained in these function words. Because "function words are hardest to recognize when they occur in clusters, such as is the, that are, and of the", Lee et al. identified a set of 12 such phrases, modified the pronunciations of these phrases according to phonological rules, and modeled the phones in them separately. The new system also incorporates "generalized triphone models", created from triphone models using a clustering procedure. One benefit of this procedure is that it provides an "ideal means for finding the equilibrium between trainability and sensitivity", which is important in view of the limitations on the amount of available training material in the Resource Management Database. CMU found that "generalized triphones outperformed triphones, while saving 60% memory". Like others reporting results at this meeting (e.g., Paul at MIT Lincoln Laboratory (2) and Weintraub et al. at SRI (4)) CMU's recent work included procedures to account for between-word coarticulation. Because the number of triphone models grows sharply when between-word triphones are considered, CMU clustered a set of 7057 triphones (that was generated from an original set of 2381 within-word triphones by considering the between-word triphones) into 1000 generalized triphone models. More complex connections are then needed to link adjacent words together in the sentence model, and the recognition algorithm must be modified. "Corrective training" [introduced by Bahl et al.] was included in the CMU speaker-independent system by using cross validation procedures and a combination of a dynamic programming algorithm to align reference sentences with misrecognized sentences in the cross-recognized training set to produce an ordered list of likely phrase substitutions. This list of phrase substitutions was then used to randomly hypothesize near-miss sentences for reinforcement learning, improving correct words and suppressing near-misses.

The SRI speaker-independent continuous speech, large vocabulary speech recognition system, DECIPHER, integrates "speech and linguistic knowledge into the HMM framework". Phonological modeling is explicitly accounted for by developing phonological rule sets based on measures of coverage and overcoverage of a database of pronunciations in order to maximize the coverage of pronunciations observed in a corpus, while minimizing the size of the pronunciation networks. The DECIPHER system incorporates probabilities into the network of word pronunciations. A number of different lexicons were studied, including those used at BBN and CMU. For the SRI lexicons studied, the mean number of pronunciations per word ranged from 1.0 to 4.2. The studies showed that careful design of the dictionary of pronunciations can yield performance improvements (i.e., automatically deriving a dictionary of most common pronunciations proved superior to the case for a dictionary carefully designed by hand by an expert linguist). Modelling a small number of multiple probabilistic pronunciations (e.g., a mean number of 1.3 pronunciations per word) showed greater performance improvements than for the case of a larger number (e.g., 4.3 pronunciations per word), perhaps because in the latter case the pronunciation networks become "too bushy". SRI attributes the success of their approach to modelling multiple pronunciations to the incorporation of constraints on the pronunciation networks. SRI's system also incorporated consideration of coarticulatory effects across word boundaries. In this implementation, "modeling acoustic variations across words was limited to initial and final phones in words with sufficient training data" (i.e., "provided that 15 occurrences of a (previous/next) phone occurred in the training database").

In contrast to the other systems described in this Session, all of which are HMM systems, the MIT SUMMIT System (5) is a phonetically based system that "attempts to express the speech knowledge within a formal framework using well-defined mathematical tools. Features and decision strategies are discovered and trained automatically, using a large body of speech data." (5) At this time, the SUMMIT system does not explicitly make use of context-dependent models. Zue et al. (5) note that the SUMMIT system, making use of 75 phoneme models, might be compared to an early version of the CMU SPHINX system that "achieved a word recognition rate of 84% and 93% using 48 and 1,000 models" [on an earlier test set]. Zue et al. go on to note that their "result of 87% on 75 models is quite competitive using a very different approach to speech recognition than hidden Markov modelling."

Preliminary results presented by Pieraccini of CSELT on behalf of the group at AT&T Bell Laboratories described a series of recognition experiments to determine the extent to which standard modelling techniques for continuous density hidden Markov models could be applied. AT&T's studies included the effects of sampling rate (downsampling to rates of 6.67 kHz and 8 kHz), and frame shifts of 15 and 10 msec. In general, superior results occurred for the higher sampling rate due to the larger bandwidth of the signal, while the shorter frame shifts produce better temporal resolution at the acoustic level.

Picheny from the IBM Watson Laboratories presented some informal results on two large-vocabulary continuous speech tasks: the 5,000 word office correspondence task and the DARPA [991 word] Resource Management task.

SUMMARY DARPA-SITE BENCHMARK TEST RESULTS

In Tables 1. and 2., the quantity "Corr" is the percentage of words in the reference string that are correctly recognized in the system's output hypothesis strings. "Sub" is the percentage of words resulting in substitution errors, and "Del" and "Ins" are the percentages of words resulting in deletion or insertion errors, respectively. The total word error percentage, "Err" = "Sub" + "Del" + "Ins". (The term "Word Accuracy" [%] is sometimes used to refer to the quantity 100 – "Err".) "Sent Err" refers to the percentage of sentences in the test material that are recognized without errors of any kind.

The data presented in Tables 1 and 2 are derived from implementation of the DARPA NIST (NBS) standard scoring software package (6) on results reported to NIST. Although the scoring software provides a great deal of data, it was thought desirable to provide a concise summary of results for the February 1989 Benchmark Tests using a consistent format, as in these tables. Test results are shown for two conditions: (1) the "word-pair grammar", a non-probabilistic list of allowable word pairs, and (2) "no grammar", in which all words are treated as equally probable.

Because differences between the results for different systems are small, the statistical significance to be attributed to these differences is not known at present. Future modifications to the scoring software may incorporate provisions for statistical tests such as McNemar's and the Cochran Q-test (7) so that significances may be assigned to these differences.

For the speaker-independent systems, results are grouped according to the amount of system training material used in preparing for these tests.

ACKNOWLEDGEMENTS

At NIST, John Garofolo was responsible for preparation and distribution of the test material and implementation of the NIST standardized scoring software for the Benchmark Tests. It is to his credit that this task was completed in a very timely manner, despite a number of hardware and software problems.

The discussion of system attributes presented in the "Overview" portion of this paper was prepared by the author using material in references 1-5. Any inaccuracy is solely my responsibility: the reader is referred to these references for more details.

REFERENCES

1. Lee, K-F., Hon, H-W., and Hwang, M-Y., "Recent Progress in the Sphinx Speech Recognition System", in Proceedings of the February 1989 DARPA Speech and Natural Language Workshop Philadelphia, February 21-23, 1989.

2. Paul, D.B., "The Lincoln Continuous Speech Recognition System: Recent Developments and Results", in Proceedings of the February 1989 DARPA Speech and Natural Language Workshop, Philadelphia, February 21-23, 1989.

3. Schwartz, R. et al., "Robust Smoothing Methods for Discrete Hidden Markov Models", to appear in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Glasgow, Scotland, May 23-26, 1989.

4. Weintraub, M. et al., "Linguistic Constraints in Hidden Markov Model Based Speech Recognition", in Proceedings of the February 1989 DARPA Speech and Natural Language Workshop, Philadelphia, February 21-23, 1989.

5. Zue, V. et al., "The MIT SUMMIT Speech Recognition System: A Progress Report", in Proceedings of the February 1989 DARPA Speech and Natural Language Workshop, Philadelphia, February 21-23, 1989.

6. Pallett, D.S., "February 1989 DARPA Speech Recognition Resource Management Benchmark Tests, Amended January 23, 1989", notes outlining February 1989 DARPA Benchmark test procedures [distributed

at the February 1989 DARPA Speech and Natural Language Workshop, Philadelphia, February 21-23, 1989].

7. Gillick, L. and Cox, S.J., "Some Statistical Issues in the Comparison of Speech Recognition Algorithms", to appear in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Glasgow, Scotland, May 23-26, 1989.

Table 1. SPEAKER-DEPENDENT TEST SET

Average results for 300 sentence utterances (12 speakers)

a. Word-Pair Grammar

						Sent
	Corr	Sub	Del	Ins	Err	Err
BBN Triphone	97.5	2.0	0.5	0.6	3.1	21.0
LL Dependent	96.8	2.7	0.5	1.0	4.2	28.0
LL Independent (109)	90.7	6.8	2.5	2.1	11.4	47.7

b. No Grammar

						Sent
	Corr	Sub	Del	Ins	Err	Err
BBN Triphone	87.0	10.0	3.0	0.8	13.8	66.0
LL Dependent	89.6	8.4	1.9	2.8	13.2	60.3
LL Independent (109)	72.6	20.5	6.9	3.8	31.2	87.0

Table 2. SPEAKER–INDEPENDENT TEST SET

Average Results over 300 sentence utterances (10 speakers)

a. Word–Pair Grammar

Systems trained on 109 speakers						Sent
	Corr	Sub	Del	Ins	Err	Err
CMU (109)	94.5	4.4	1.1	0.6	6.1	34.3
SRI (109)	91.6	5.9	2.5	0.4	8.8	42.7
LL (109)	91.3	6.4	2.2	1.2	9.8	44.3
Systems trained on 72 speakers						
SRI (72)	91.1	6.4	2.5	0.3	9.2	43.7
LL (72)	90.3	6.8	2.9	1.3	11.0	50.0
MIT (72)	87.6	10.3	2.1	1.2	13.6	54.7

b. No Grammar

Systems trained on 109 speakers						Sent
	Corr	Sub	Del	Ins	Err	Err
CMU (109)	80.2	17.4	2.4	4.8	24.5	76.7
SRI (109)	73.8	21.0	5.2	1.5	27.7	87.0
LL (109)	75.3	18.7	6.1	3.1	27.9	82.0
Systems trained on 72 speakers						
SRI (72)	70.6	23.4	6.1	1.8	31.3	87.7
LL (72)	72.3	21.0	6.7	3.0	30.7	86. 3
MIT (72)	49.8	40.0	10.3	3.5	53.8	97.4